An Analytical Review on Feature Reduction for Big Data Analytics using Machine Learning

Rachna Kulhare, S Veenadhari, Neha Sharma

Department of CSE, RNTU, Bhopal (M.P), India

Corresponding author: Rachna Kulhare, Email: rkulhare123@gmail.com

There is an explosive growth of data due to advancements in computer methods. With ML techniques, working on a really massive quantity of data is a big problem. As a result, handling and computing on a very vast, varied, and diverse dataset seems a difficult undertaking. The purpose of this study is to provide a quick overview of several dimensionality reduction/feature selection techniques. A summary of the contributions of scholars to the development of feature selection methods for huge datasets is provided. This study is driven to create a hybrid, resilient, adjustable, as well as dynamical feature selection approach to classify huge datasets by examining current challenges.

Keywords: Machine Learning, Feature selection, Data Mining, Large Datasets, Dimensionality Reduction.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 525–534. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-54

1 Introduction

As technology advanced, the amounts of information grew exponentially, finding it difficult to use certain information in just about any field without proper pre-processing, resulting in increased complexity and precision issues, which eventually led to a slew of issues, including processed, analyzed, highly classified, and secrecy [1]. For a maximum of the latest statistics mining and ML procedures, learners from these massive datasets are a big challenge. Big data can therefore seem easy to manage, but due to its characteristics' sophistication, heterogeneity, and hybridity, it actually demands a rather complex method. This method is called the process of exploration of knowledge:

- Data recording: Involves various problems and tools related to data collection and storage.
- Pre-processing of data entails all cleaning procedures and the conversion of collected data into an analysis-ready format to optimize the analytical phase.
- Data analysis: This work entails analyzing data using a variety of logical reasoning algorithms to explore each segment of the data in order to provide useful findings.
- Data visualization and interpretation: This stage entails the effective display of data using a variety of approaches in order to analyze the significance and meaning of the findings [2].

Due to the obvious large amounts of raw data that are kept, it really has gotten a lot of interest in many fields, including biotechnology, health, advertising, and financial services. The revamping of methods and their incorporation in parallel settings may be designed to change machine learning techniques for large data issues [3]. Feature selection, which is often used to discover correlated features and eliminate redundant or uncorrelated data from the feature collection, is among the most important data analysis methods. Random or noisy qualities can make it difficult for a classifier to create reliable connections and duplicated or correlated features can confuse a classifier without adding any value. [4][5]. The primary principle behind feature selection is to invent a fresh dataset with neither redundant nor unnecessary characteristics nor the existing data pattern while retaining every one of the existing dataset's important information. With their capacity to reduce dimension, feature selection techniques are already extensively used, and they demonstrate their value when the number of features is great, but the number of observations is limited. As opposed to feature extraction, feature selection tries to discover features that can accurately and briefly explain the original dataset, whilst the converse aims to produce innovative apps following the original information. It should always be mentioned that certain linked aspects may be superfluous due to the fact that there might be some more than one.

The following are some of the benefits of making feature selection on a data set:

- The selected features may be used to create a short model for explaining original data, which is good for increasing performance.
- The chosen characteristics can represent the essential qualities of the original data, making them useful for reliably tracing data expression.
- The selected characteristics can assist the decision-maker in sorting through a huge amount of data.
- It allows the machine learning system to understand more quickly.
- It simplifies a model's complexities and makes it easier to understand.
- If the appropriate subset is chosen, it increases the accuracy of the model and decreases overfitting.

2 Big Data Processing: An Overview

With the advancement of computer technology, the volume of data has increased exponentially, making it difficult to use such data without proper pre-processing. This, in turn, leads to greater complexity and truthfulness, resulting in a slew of issues, including storage, analysis, security, and privacy. As a result, big data might appear to be simple to manage, as seen in [6]-[11]:

Volume

- The size of data is big — it must be divided into manageable pieces.
- Data must be managed in different systems in parallel.
- Data must be processed concurrently through several programmed modules.
- Data must be stored and processed once for volume completion.
- Data must be processed from every failure point, as it is incredibly large that the process is restarted from the start.

Velocity

- Data must be processed during data processing at streaming speeds.
- For multiple acquisition points, data must be processed.

Variety

- It is important to process data from various formats.
- It is important to process data of various types.
- It is important to process data from various systems.
- It is important to process data from different regions.
- Therefore, the sophistication of big data requires the use of several algorithms for the rapid and efficient processing of data. Multipass processing is important for many types of data and scalability. The processing of large data involves the simplest management of an extremely high-performance computing environment that allows linear scalability to be modified to achieve performance.

3 Types of Feature Selection Techniques

In broad data processing, the key challenge is the development of a solid pre-processing model. It is obligatory to face these obstacles so that only feature selection strategies' most important features will minimize volumes and complexity.The pre-processing stage is the basis for the accuracy of the analysis. The data creates many challenges, such as enormous complexity and multiple functions and attributes, even with small bases. An adequate processing phase for high-quality analysis is critical and essential. One of thepre-processing phase aims to reduce a dataset’s dimensionality and sophistication, which is done by selecting features. Most of the feature selection strategies are as follows:

Filter Methods

Filter techniques are an independent pre-processing step from subsequent learning algorithms. To pick characteristics, they use separate techniques. In order to determine the degree of significance of each characteristic to a target variable, a set of characteristics is selected by an assessment criterion or a ranking.

Wrappers Methods

Wrappers are feature selection strategies that are used to evaluate the accuracy of a prediction model trained on a subset of attributes. The evaluation is performed with the help of a classifier that calculates the relevance of a subset of characteristics. This approach has proven to be effective, and while it is computer-intensive, thus it is not widely used.

Embedded Methods

Combine filter characteristics and methods of packaging. As the Filter methods have proved less effective but faster, and the Wrapper methods are more efficient but particularly computer-efficient with large datasets, a solution combining the pros of the two types is necessary.

Hybrid Methods

Many methods of primary conjuncture collection accompany this method.

Ensemble Methods

It uses a combination of features of various basic classifications. It requires the use of various subsets of functions.

Integrative Methods

For feature selection, incorporate external information. A significant number of classification problems employ function selection in order to make classifications more precise and reduce time. There can be a great deal of noise in a full set of functions. With less detail, the data is summarized or represented. This is useful for displaying dimensional data or simplifying knowledge that can then be used in a supervised learning method. For classification and regression, several of these methods of reduction are relevant. Principal Component Analysis (PCA), Linear Discrimination Analysis (LDA), Autoencoders, etc., are the most common algorithms: Some of the extraction methods of features are discussed below.

Linear Discriminant Analysis (LDA): In being uncorrelated and optimized class separation, a set m of linear connections (discriminating functions) of n input characteristics, with x y, are generated. The dataset's new basis has become these discriminating functions. The linear discriminate functions are used to estimate every quantitative component throughout the information, converting the dataset from y-dimensionality to x-dimensionality. The target column must be defined initially before using the LDA method for dimensionality reduction. The overall number of decreased measures m is equal towards the target column's number of classes minus one, or, if less, the information's amount of numerical columns.

Autoencoder: The autoencoder is a NN (neural network) with at least one hidden layer and an x y total of n output nodes that have already been trained to reproduce the input vector just on the output layer by using the backpropagation approach. The numeric columns within the information are decreased by symbolizing the input vector with the output of the hidden layer. The encoder was called out from input towards the hidden layer of m units in the initial section of the autoencoder. The data set's y dimensions are compressed into x-dimensional space. The decoder first from the secret layer to the output layer is the auto encoder's second component. The decoder extended the data vector with an x-dimensional space into another original YD dataset and restored the information towards its actual numbers.

T-distributed Stochastic Neighbour Embedding (t-SNE): Depending on nonlinear local connections among pieces of data, the dataset's n numerical columns were decreased to fewer x (x <y). In order to model identical items at near locations and distinct things in the special low multidimensional region at distant places, every high-dimensional item is modeled using a double or three-dominated spot.

The data points will be modeled using a multivariate normal numerical column distribution during the first phase. The next step is to substitute this probability with a smaller t-distribution that closely resembles the initial multivariable standard.

The t-distribution allows you to choose any location throughout the data set to be a neighbor within lower-dimensional space. Its data density is controlled by the confusion parameter, which is defined as that of the efficient number of neighbors for each location. The larger the global scope of both the data, the more confusion. Just the existing dataset is compatible only with the t-SNE method. The model can't be exported to use with new data.

Missing Values Ratio: Information columns with those kinds of null values are unlikely to provide any relevant information. Information columns may indeed be removed with such a lost price ratio greater than a certain threshold. The steeper the drop, the greater the threshold.

Low Variance Filter: There really is no understanding of information columns with little data modification, comparable to both the prior technique. As a result, all data columns underneath the threshold can be removed. This is a viable option. Because the variance is dependent on the columns array, normalization is compulsory previously using this approach.

High Correlation Filter: Data columns having predictable styles are likely to contain comparable data, and merely one of them would be enough to classify them. The Pearson Product Moment's coefficient of correlation b/w the numerical columns and the Pearson's pickup parameters between the numerical columns are calculated here. Again for the final prediction, only one column is kept for every pair of columns with such a fairly correlation greater than that level. Because correlation is dependent just on the column set, standardization is necessary before using this approach.

Random Forests/Ensemble Trees: Decision-taking tree algorithms, often based on random forests, are beneficial for column selection in addition to becoming fast classifiers. To predict the target classes, a large and complex collection of trees is built, and afterward the characteristics for each column are used to determine another very informative subset of columns. Every tree is developed in a small percentage of the total number of columns, resulting in a huge number of extremely shallow trees. Considering columns are frequently picked by way of the best split, it will almost certainly become an insightful column that we must keep. We provide a score to each column based on the number of columns split by the number of times they have become an application.

PCA (Principal Component Analysis): PCA (primary component analysis) is a statistical method that turns a dataset's original n numerical variables together into a new collection of n dimensions called main elements. The first key focus element has the largest variance conceivable due to the conversion; each subsequent primary component has the maximum impact due to restrictions that are orthogonal to (i.e., unrelated to) the previous main components. Data dimensions are decreased but the bulk of information, i.e., data variance, is maintained just by retaining the initial x is less than y key components. It's also worth noting that the theory's current parameters (PCs) aren't any actual variables anymore. When you apply PCA to your dataset, it loses its generalization ability. PCA is not really the modification for you, but if your study's findings must be interpretable.

Backward Feature Elimination: The chosen categorization algorithm is trained on n input columns in each cycle in this method. The design would then be constructed on x-1 columns after one column is eliminated. The input column, which deletion resulted throughout the least increase in mistake rate, is removed to make room for x-1 columns. After then, the categorization will be duplicated with x-2 and so on. Every k iteration generates a model trained on an x-y number of the column and an error rate e. (y). We specify the minimum number of columns necessary with the specified machine learning method to attain this classifier accuracy by setting a specified maximum error rate.

Forward Feature Construction: This is the reverse technique for the removal of functionality. We begin by simply adding a single column, which gradually increases its efficiency by adding one column at a time. In terms of time and calculation, all algorithms, backward removal, and forward functions are very costly. It is only useful if it is used on a data set with a relatively small number of columns.

Swarm Intelligence: Swarm intelligence is shown to be effective in addressing NP-hard problems. Therefore, using the swarm intelligence technique would result in a set of features that can satisfy the condition called a fitness function that can be further used in algorithms for machine learning. The Swarm-based feature selection or dimension reduction technique searches for functions in the search space based on a search strategy such as a global or local search process. There are some basic stages described in each swarm intelligence technique as below:

Imitation of the estimation parameters and swarm population.
Iteration termination conditions are determined.
Assess the fitness of each population.
Find the right solution locally.
Population Position Updates.
Return to the best global solution.

Some of the best techniques used for swarm-based dimension and optimization work in recent research are: particle swarm optimization (PSO), ant colony optimization (ACO), grass-hopper optimization (GHO), Crow Search Optimization (CSO), grey wolf optimization (GWO), etc.

4 Related Work

Siddiqi et al. [12] suggested the architecture of the genetic algorithm (GA) wherein the fitness of solutions is made of two components. A feature-selection measure will be the first, like MI, JMI, or mRMR. The second term is the overlap factor, which is accountable for GA diversity. Experimental results show that the proposed algorithm can return many high-quality solutions, which also has limited overlap. Many solutions carry considerable advantages when none or missing values include test data. Two publicly accessible time-series datasets were used for experiments.

The extracted features are also utilized to perform forecasting that uses a basic Long Short-Term Memory (LSTM) framework. The solution output of predicting employing distinct feature sets is evaluated.

Kong et al. [13] presented a decentralized fuzzy rough set (DFRS)-based feature selection method that divides and allocates jobs to numerous parallel computing nodes. The key problem is to keep global information about each dispersed node without having to keep track of the full fuzzy connection matrix.

Ding et al. [14] proposed a multiple-specific feature ensemble selection (MRFES) approach that relies on multilayer coevolutionary consensus MapReduce (MCCM). An accurate MCCM framework is

formulated and solved to sustain the set of large-scale feature ensemble sets of data with method adds feature sources and to investigate the unified aggregation of consistency among locally and globally supremacy solutions accomplished by co-evolutionary memeplexes associated in the cooperative feature ensemble selection procedure.

Using grey wolf optimization and particle swarm optimization, Hasnony et al. [15] suggested binary variants of the wrapper function selection. The K-nearest neighbor classifier with Euclidean separation matrices is used to find the best solutions. A messy tent map helps to prevent the algorithm from being locked to the dilemma of local optima. The sigmoid function converts the search space from a continuous vector to a binary one to be sufficient for the choice of feature problem.

A dynamic function mask for clustering high dimensional data streams was proposed by Fahy et al. [16]. Redundant features are masked, and the unmasked, important features are clustered. The mask is changed accordingly if the perceived value of a feature changes; previously unimportant characteristics are unmasked, and features that lose significance are masked.Yang et al. [17] proposed an instructive functional clustering and selection technique for selecting relevant and diverse genes through gene expression data. There are two steps to this phase. In the first stage, a feature clustering (FC) approach divides total genes into multiple gene groups. In FC, a set of feature weights is generated to recognize the relevance of every gene and group the genes into distinct gene clusters depending on the feature weights. A stratified feature selection (SFS) approach is utilized in phase 2 to choose genes from various gene clusters and combine them to make a complete feature set.Liang et al. [18] presented a hybrid genetic algorithm called the wrapper-embedded function strategy to select (HGAWE) that mixes evolutionary algorithms (global search) with embedded regularization methods (local search). Zaffar et al. [19] conducted a study of feature selection algorithms' effectiveness using students' data sets. The results of the different FS techniques and classifiers might also aid research in the future in determining the optimal combinations of FS methods and classifiers. Because educational stakeholders must make decisions based on the findings of prediction models, selecting relevant characteristics for the student prediction model is a delicate challenge.Fong et al. [20] proposed a lightweight feature collection. The features are specifically developed for on-the-fly extraction of data streams utilizing the swarm searching style Accelerated Particle Swarm Optimization (APSO), improving analytical accuracy and reducing process time. A collection of big data with an unusually high degree of dimensionality is analyzed in this study for performance measurement.Lin et al. [21] proposed an optimization method in which the weight within each mark's feature rankings and the resulting feature rank list are defined as two variables involved. The purpose is to minimize the weighted total divergence between both the resulting feature rank list and each mark's feature rank list.Peralta et al. [22] presented an evolutionary computation-based feature selection technique that leverages the MapReduce paradigm to extract subsets of features from huge datasets. The method breaks down the entire data into the frames of instances in order to understand by them in the map phase; after which, inside the reduction phase, this then integrates the extracted features into the final vector of feature weights, enabling an adaptable implementation of the feature selection method to use a threshold to ascertain the selected subset of features. Three classifiers are used to assess the feature selection technique.

Table 1 gives the comparative study of various methods for feature selection methods in big data analytics. From the given study, fig 1 shows the comparative state-of-art of different machine learning approaches for big data feature selection or reduction. From the figure, it can be analyzed that fuzzy logic and evolutionary algorithm achieved better performance. So, in the future, researchers can adapt these algorithms to improve their efficiency of work.

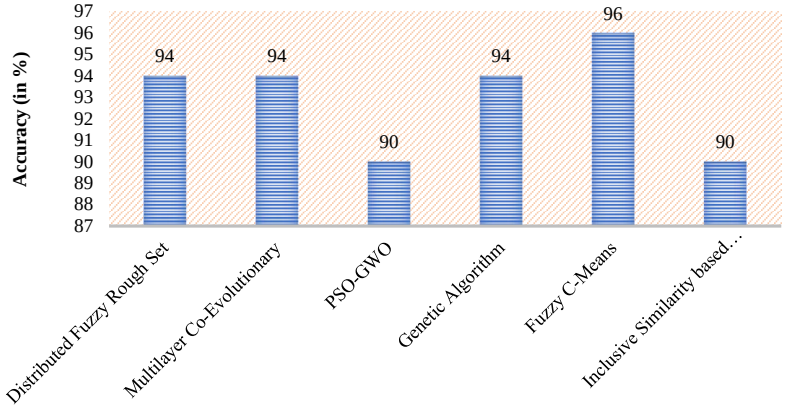

Fig. 1. Comparative State-of-art of ML for Feature Selection and Reduction

Table 1. Some Exiting Feature Reduction Techniques

Ref	Technique Used	Discussions
[12]	Genetic Algorithm	Give diverse and accurate results when tested on a large dataset. But computational time is more.
[13]	Distributed Fuzzy Rough Set	The fuzzy logic achieved 94% of accuracy but faced issues when datasize was increased randomly.
[14]	Multilayer Co-Evolutionary	The hybrid approach achieved 94% of accuracy, but it was observed that the performance degrades with an increased level of the noise level in data.
[15]	PSO-GWO	The hybrid approach achieved 90% of accuracy, but the complex architecture archives an overfitting problem.
[18]	Genetic Algorithm	The algorithm achieved 94% of accuracy but faced high computational time.
[20]	Particle swarm optimization	The performance was good, but it faced high computational time.
[23]	Fuzzy C-Means	The performance of FCM was 96% of accuracy and faced high computational time.
[24]	Canonical PSO	It shows its efficacy over PSO but faces high computational time.
[25]	MO-PSO	The error rate was 0.5 while reduction, but the increasing iteration of the hybrid approach decreases the speed.
	MOGA	The error rate was 0.47,duringthe reduction
	MOCSO	The error rate was 0.01,duringthe reduction
[26]	Shared Nearest Neighbor clustering	Performance was good enough, but redundancy was achieved.
[27]	Inclusive Similarity-based clustering	The accuracy was 90%, but time consumption was very high.

[28]	GWO	Performs better than PSO and GA but doesn't handle multi-objective tasks.

5 Conclusion

The overfitting issue of classifiers for big databases will be reduced by using Dimensionality reduction as well as feature selection strategies. This work focuses on providing an analytical evaluation of existing research issues for feature reduction on huge data sets. In comparison to standard feature selection approaches, bio-inspired algorithms such as swarm intelligence and genetic algorithm, and so on. are common way for locating relevant characteristics, according to the literature review. Along with that, this paper also presented an accurate comparison of some existing machine learning algorithms for feature selection and reduction. From the results, it has been seen that most of the researchers had contributed their efforts in designing swarm intelligence or evolutionary optimization algorithms. But from the results, it can be observed that evolutionary optimization algorithms can outperform better. So, in the future, researchers can focus their efforts on designing a hybrid optimization algorithm for feature reduction for both classifications as well as time-series problems.

References

[1] Tuo, Q., Zhao, H. and Hu, Q. (2019). Hierarchical feature selection with subtree based graph regularization. *Knowledge-Based Systems*, 163: 996–1008.

[2] Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems*, 86: 33–45.

[3] Ji, B. et al. (2020). Bio-Inspired Feature Selection: An Improved Binary Particle Swarm Optimization Approach. *IEEE Access*, 8: 85989-86002.

[4] Elhariri, E., El-Bendary, N. and Taie, S. A. (2020). Using Hybrid Filter-Wrapper Feature Selection with Multi-Objective Improved-Salp Optimization for Crack Severity Recognition. *IEEE Access*, 8:84290–84315.

[5] Larabi-Marie-Sainte, Souad. (2021). Outlier Detection Based Feature Selection Exploiting Bio-Inspired Optimization Algorithms. *Applied Sciences*, 11(15): 6769.

[6] Philip, C. C. L. and Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences (Ny)*, 275: 314–347.

[7] Shahmoradi, M. R. et al. (2019). Multilayer overlapping community detection using multi-objective optimization. *Future Generation Computer System*, 101: 221–235.

[8] Landset, S. et al. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2 (1): 1–36.

[9] Kushmerick, N., Weld, D.S. and Doorenbos, R. (1997). Wrapper Induction for Information Extraction. *PhD Thesis. https://dada.cs.washington.edu/research/tr/1997/11/UW-CSE-97-11-04.pdf*

[10] Naseriparsa, M., Bidgoli, A. M. and Varaee, T. (2014). A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms. *arXiv*.1403.2372, 69 (17): 28–35.

[11] Tsymbal, A., Pechenizkiy, M. and Cunningham, P. (2005). Diversity in search strategies for ensemble feature selection. *Information Fusion*, 6(1): 83–98.

[12] Siddiqi, U.F., Sait, S.M. and Kaynak, O. (2020). Genetic algorithm for the mutual information-based feature selection in univariate time series data. *IEEE Access*, 8: 9597–9609.

[13] Kong, L. et al. (2020). Distributed Feature Selection for Big Data Using Fuzzy Rough Sets. *IEEE Transaction & Fuzzy Syst*em, 28(5): 846–857.

[14] Ding, W., Lin, C. and Pedrycz, W. (2018). Multiple Relevant Feature Ensemble Selection Based on Multilayer Co-Evolutionary Consensus MapReduce. In *IEEE Transactions on Cybernetics*, 50(2): 425-439.

[15] El-Hasnony, I.M. et al. (2020). Improved Feature Selection Model for Big Data Analytics. *IEEE Access*, 8: 66989–67004.

[16] Fahy, C. and Yang, S. (2019). Dynamic Feature Selection for Clustering High Dimensional Data Streams. *IEEE Access*, 7: 127128–127140.

[17] Yang, Y. et al. (2019). Informative Feature Clustering and Selection for Gene Expression Data. *IEEE Access*, 7: 169174-169184.

[18] Liu, X. et al. (2018). A Hybrid Genetic Algorithm with Wrapper-Embedded Approaches for Feature Selection. *IEEE Access*, 6: 22863-22874.

[19] Zaffar, M., Hashmani, M.A. and Savita, K.S. (2017). Performance analysis of feature selection algorithm for educational data mining. In *IEEE Conference on Big Data and Analytics*, 7-12.

[20] Fong, S., Wong, R., and Vasilako, A. (2016). Accelerated PSO swarm search feature selection for data stream mining big data. *IEEE Transactions on Services Computing*, 9(1): 33–45.

[21] Lin, Y. et al. (2016). Multi-label feature selection with streaming labels. *Information Sciences (Ny)*, 372: 256–275.

[22] Peralta, D. et al. (2015). Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach. *Mathematical Problems in Engineering*, 2015.

[23] Khan, M. A. et al. (2021). An Integrated Design of Fuzzy C-Means and NCA-Based Multi-properties Feature Reduction for Brain Tumor Recognition. *Signal and Image Processing Techniques for the Development of Intelligent Healthcare System*, 1–28.

[24] Gu, S., Cheng, R. and Jin, Y. (2016). Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Computing*, 22(3): 811–822 .

[25] Yan, D. et al. (2020). Single-Objective/Multiobjective Cat Swarm Optimization Clustering Analysis for Data Partition. *IEEE Transactions on Automation Science and Engineering*, 17(3): 1633-1646.

[26] Wang, S. and Eick, C. F. (2017). MR-SNN: Design of parallel Shared Nearest Neighbor clustering algorithm using MapReduce. In *IEEE International Conference on Big Data Analysis*, 312-315.

[27] Sangeetha, J. and Prakash, V. S. J. (2017). An Efficient Inclusive Similarity Based Clustering (ISC) Algorithm for Big Data. In *World Congress on Computing and Communication Technologies*, 84-88.

[28] Emary, E. et al. (2015). Feature subset selection approach by gray-wolf optimization. In *Afro-European conference for industrial advancement*, 1–13.